NOAA NESDIS
Central Satellite Data Processing Center

Comprehensive Large Array-data Steward-ship System (CLASS) Concept of Operations
For the NPOESS Preparatory Project (NPP) Archive and Distribution Segment (ADS)

CSC

July 20, 2001

Computer Sciences Corporation
Laurel, MD

TABLE OF CONTENTS

TABLE OF FIGURES

1.0 Overview

The National Polar-Orbiting Environmental Satellite Systems (NPOES) Preparatory Project (NPP) Archive and Distribution Segment (ADS) architectural concept is based on Comprehensive Large Array-data Stewardship System (CLASS) architecture and encompasses the major functions associated with the ingest/inventory, processing, archive/storage, retrieval, and distribution of data and derived data products from the NPP polar-orbiting sensors. The architecture relies on a series of integrated process management elements, each of which satisfies a collection of functional requirements. NOTE: To stay in line with other NPP draft documents, the following are functional requirements describing "what" the concept of operations is to be, not the "how." The major system elements are:

System Management -

> Directs and coordinates the interaction of all other system elements
>
> Controls the security and integrity of the ADS
>
> Controls/integrates all hardware components of the ADS, including those which exist at initial implementation and all life cycle upgrades
>
> Controls/integrates all software components of the ADS
>
> Controls/integrates the communication among elements and between the ADS and external (to the ADS) systems
>
> Controls/coordinates all scheduled tasks associated with ADS operations
>
> Generates metadata associated with hardware, software, and telecommunications tasks

Metadata Management -

> Coordinates the ingest, storage, and distribution of information relating to the ADS operations (logging, configuration schematics, metrics, etc.)
>
> Coordinates the ingest, storage, and distribution of information relating to the NPP data and metadata being ingested, stored, retrieved, and/or distributed

Ingest Management -

Coordinates/quality controls the acquisition and integrity of data, derived data products, and information from NPP suppliers

Creates inventory metadata, associated with acquired data, data assessments, associated ancillary data, science algorithms, software, documentation, calibration parameters, other metadata, and browse imagery if required and feasible

Supplies data and/or derived data products for standing orders to the data and product retrieval and distribution element

Data and Product Archive/Preservation and Migration Management -

Coordinates/maintains the internal storage of data on media, including magnetic disk, optical media and/or tape media

Creates metadata associated with stored data, including physical location and media type via a storage management system

Ensures media is exercised on a periodic basis

Manages the migration of hardware refreshment and media refreshment, always maintaining the quality/integrity of the data

Customer Interface Management -

Provides all interfaces through which customers may gain access to publicly available processors and elements

Provides mechanisms by which customers may interact with ADS metadata databases, including inventories, browse imagery, and system information

Provides e-commerce tools with which customers may select data, data products, sub-setted and super-setted data for display and/or retrieval and, if appropriate, purchase

Data and Product Retrieval and Distribution Management -

Coordinates/quality controls the extraction of data, and associated ancillary data, science algorithms, software, documentation, calibration parameters and other metadata from storage media

Manages the placement of extracted data on appropriate media, including both on-

line storage and long term media

Coordinates/controls the distribution of data, and associated ancillary data, science algorithms, software, documentation, calibration parameters and other metadata for standing orders (subscriptions) and retrospective (on-line and off-line) orders
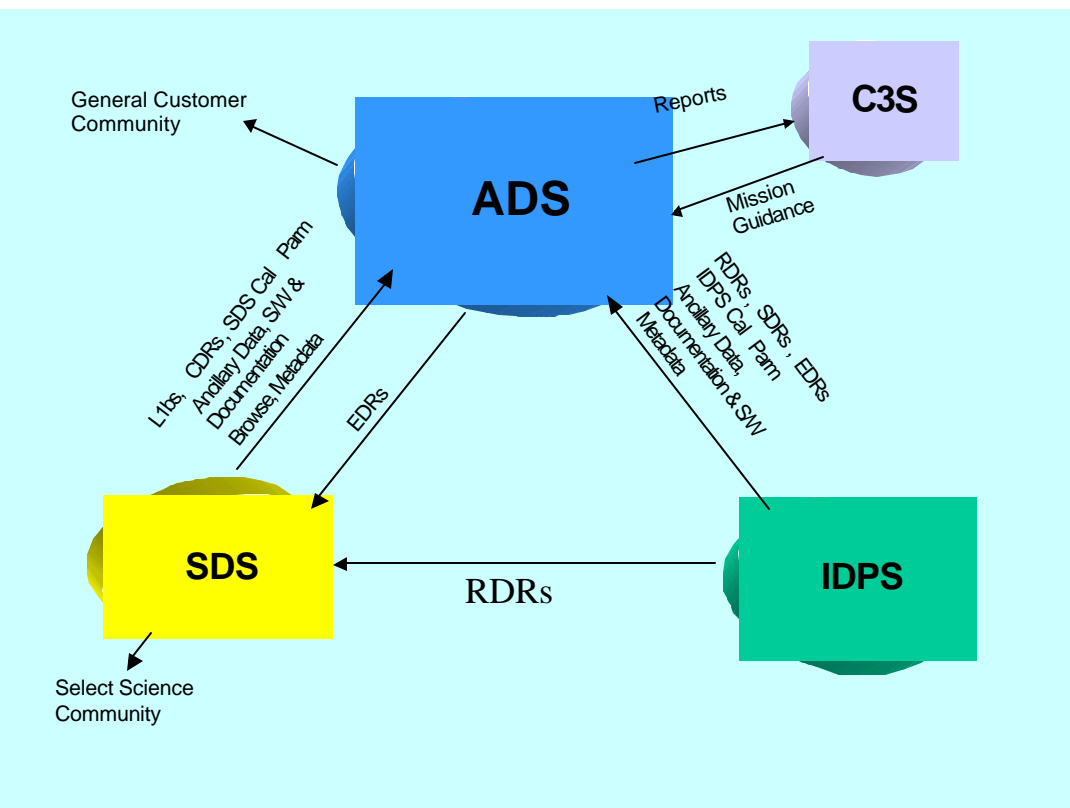
Figure 1 shows the data streams that are expected from the other NPP segments and the ones expected to be supplied to the other NPP segments from the ADS:

Figure 1 Shows the relationship of the ADS with the IDPS and SDS _and C3S._

Raw Data Record (RDR)	Sensor Data Record (SDR)
Environmental Data Record (EDR)	Calibration Parameters
Ancillary Data	Documentation
Software (S/W)	Metadata
Level 1b (L1b)	Climate Data Record (CDR)
Browse images	Mission Guidance
Status Reports	

Figure 2 depicts the major system elements as described above.
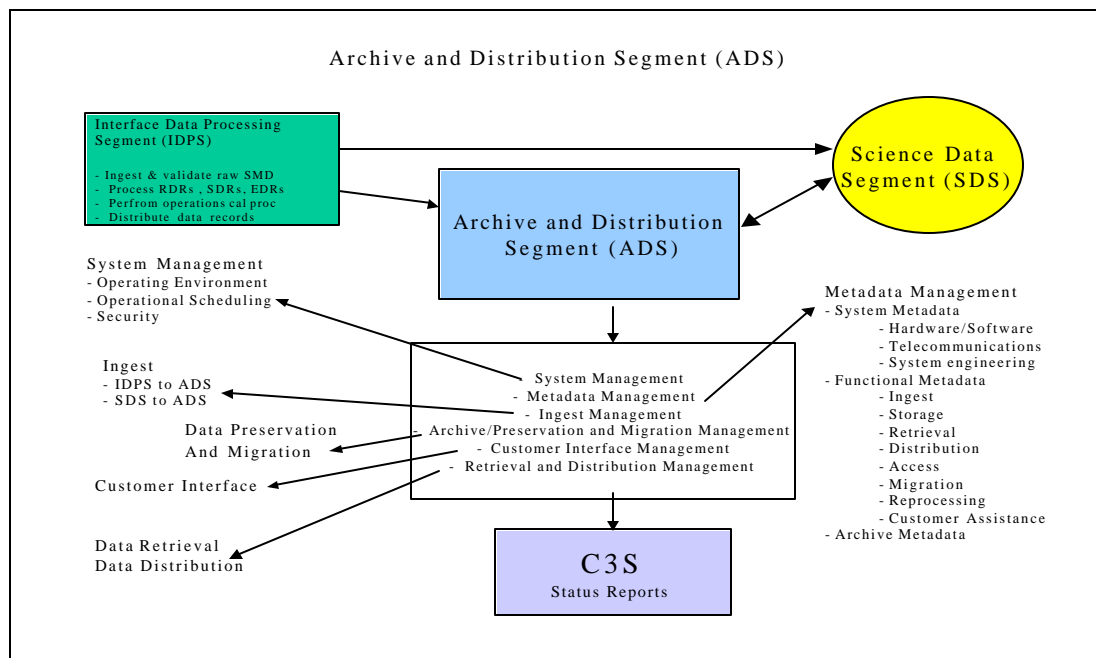

Figure 2 Major System Elements & Sub-elements

The concepts outlined in Sections 2.0 through 8.0 reflect only a portion of the overall CLASS data management responsibilities: those associated directly with the ADS and its system interactions with the IDPS, C3S and SDS.

2.0 System Management

2.1 Operating Environment

The ADS will consist of one or more open systems processors connected to external systems (input and output) via telecommunications links. The size and scale of these processors (i.e., Central Processing Unit (CPU) speed, memory capacity, cache size and disk capacity) and the external links, will be determined through analyses of NPP input, output, and processing metrics. Operational processors will have additional processing power with some redundancy, which, in addition to being used for extra processing capacity, can be initiated and used as the primary processor.

The ADS will provide and manage the operation of one or more "dedicated" telecommunications links with the Integrated Data Processing Segment (IDPS), the major supplier of NPP data, derived data products (xDRs), associated ancillary data, software, documentation, IDPS calibration parameters, and IDPS-related metadata. The IDPS-to-ADS link will, at a minimum, be maintained as a twenty-four hour by seven-day (24x7) operation with a TBD availability time. The IDPS-to-ADS link will be sized with sufficient bandwidth for initial ingest of data, to allow for retransmission or re-ingest of erroneous or garbled data, and/or allow for scheduled and unscheduled hardware maintenance tasks to be performed.

The ADS will utilize one or more telecommunications links with the NPP Science Data Segment (SDS), the major supplier of climate data records (CDRs), reprocessed data and Level 1b data products. The SDS-to-ADS link will be maintained as required (TBD) to transmit high volumes of data and information either to or from the SDS. The SDS-to-ADS link is to be managed by the SDS, with use coordinated between the two segments.

The ADS will maintain public access telecommunications links for data distribution, either via the current Internet or its subsequent replacement network(s). Additionally, the ADS may be interconnected to certain customers via dedicated or shared non-public-access telecommunications links. Such customers, potentially periodic or subscription customers and/or high data volume customers, will be provided access through coordination with ADS data management agreements. Such links may be operated and maintained by the ADS, by the customer, both through agreement, or via third parties such as dedicated government or academic networks.

As required by National Oceanic and Atmospheric Administration (NOAA) security, the ADS will provide sufficient security of system elements accessible by other systems, internal and external, including Internet access. The ADS will solely operate, manage, and maintain this security measure, having potentially no impact on IDPS, C3S or SDS interface requirements.

The ADS will operate and maintain a mass-storage system capable of the continuous, ongoing

storage of incoming NPP data records, IDPS calibration parameters, documentation and software, metadata, and ancillary data as ingested from the IDPS. Additionally, the ADS will maintain a capacity to store all data, derived data products, and associated ancillary data, science algorithms, software, documentation, SDS calibration parameters and other metadata as estimated to be received from the SDS. Finally, the ADS will maintain a capacity to store all metadata created at the ADS for the provision of products and services to its customers. The actual storage media will be transparent to both suppliers and customers and may be a combination of any/all physical media exploiting current/future technologies.

Figure3 depicts the concept of the system management of the ADS. Operational data flows to the ADS through the dedicated telecommunications lines and the system manager directs the other management processes to ingest, inventory, archive and distribute the data to the customers.
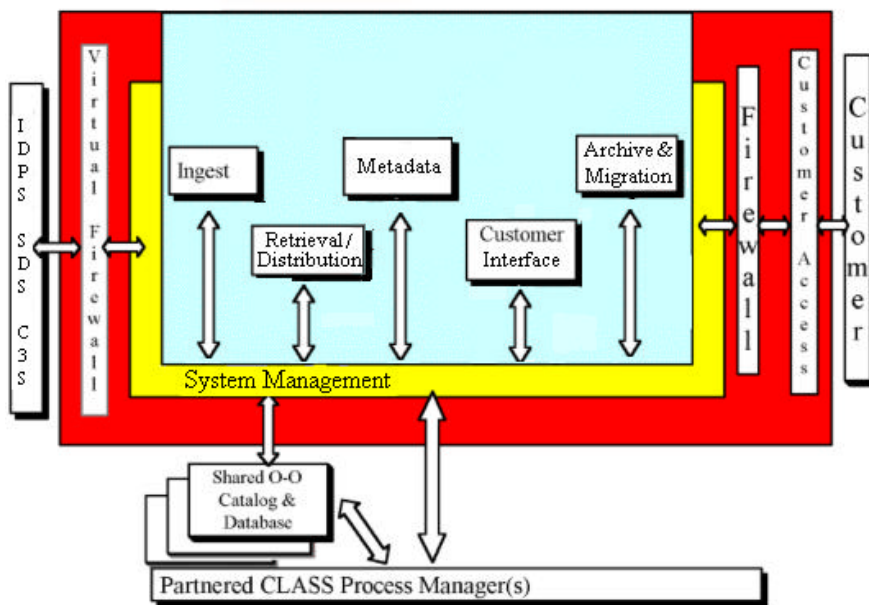
Figure 3 Operational Concept

2.2 Operational Scheduling

The ADS will operate and maintain ingest, storage, and some aspects of retrieval and distribution using a method of operational scheduling which, other than periodic, scheduled, and unscheduled maintenance tasks, will operate without human intervention. In particular, "handshakes" with IDPS/SDS processors will be maintained, providing for ingest of data, data products, and information as it becomes available. Similarly, "handshakes" if implemented for the interface with ADS customers who receive data on a periodic or subscription basis, and who have

arranged to have data "pushed" to their systems, will be maintained.

Automated logging of all system activities including data transmissions, ADS metadata generation, hardware "health," external links to customers, processing throughput, unauthorized attempts to access the ADS, etc., will be performed. This logging will be the basis of "system metadata." At a minimum activity start and end times, nature of the activity, and success/failure will be recorded. During the actual performance of an activity, the current status will be maintained.

Scheduled hardware maintenance will be performed without loss of operational functionality through the swap of secondary (additional) processors for any primary processors undergoing maintenance.

2.3 Security

System backups, including "disaster" backups of all on-line storage devices will be performed on a periodic, scheduled, automated basis. Additional processors will be maintained for ingest, storage, and retrieval functions which will automatically swap over at time of need. Sufficient on-line disk capacity will be provided on physically segregated, reserve devices, to allow continued operations in case of component failure.

Also, ADS security will be properly maintained in accordance with NOAA security regulations to prevent unauthorized access to the system components and data/information.

3.0 Metadata Management

Three areas of metadata management are considered critical to the ADS conceptual design: system internal metadata; ingest/storage/archival/retrieval/distribution management metadata information; and inventory metadata associated with the data, data products, and information being stored/archived.

3.1 System Metadata

All operational aspects of the ADS will be monitored through the creation of, storage of, and access to information about system activities and system maintenance. Wherever practical, this information will be generated and stored automatically by ADS system management functions. Information thus generated will be stored in a form and fashion, which allows rapid, indexed access; is presented in a human readable form; and will be an integral part of an ADS configuration management (CM) process. The following activities will be monitored, maintained, and made available for the life of the ADS:

- hardware operational statistics, including startup, operational, periodic, and aperiodic maintenance times

- software operational information, including revision control information, start and end of operational life cycle, test data location, test data results, and module or process documentation

- telecommunications operational information, including startup, operational, periodic, and aperiodic maintenance times, as well as periodic volume throughput metrics

- system engineering information, including configuration diagrams, hardware capacities and functionality, and system/COTS software

3.2 Functional Metadata

Information about the ongoing data management activities of the ADS will be maintained. These activities will focus on the metrics involved in ingesting, inventorying, storing, retrieving, and distributing any and all NPP data, derived data products and information. Such information will be stored in a form and fashion (inventoried) which allows rapid, indexed access; is presented in a human readable form; and will be an integral part of an ADS database management (DBM) process. The following activities and information will be monitored, maintained, and made available for the life of the ADS:

- *Ingest* including [but not limited to - BNLT] inventory of data such as: supplier (who); data/information identifier, processing algorithm and methodology used, processing date/time of ingest (what); quality checks (checksum/signature); start and end date/time of transmission (when); start and end time of record (where/when); success/failure indication (how); volume (what); and data set name [DSN] (what)

- Data and Product *Archive/Preservation and Migration* including [BNLT] storage media (disk, platter, cartridge, etc.); media identifier; checksum/signature; start/end time of storage process; DSN; and associated ancillary data, science algorithms, software, documentation, calibration parameters, other metadata and stored data volume, "old" media type and identifier; "new" media type and identifier; start/end date/time of migration; and volume migrated

- Data and Product *Retrieval and Distribution* including [BNLT] storage media; media identifier; checksum/signature; start/end time of retrieval process; DSN; and associated ancillary data, science algorithms, software, documentation, calibration parameters, other metadata and retrieved data volume, start/end date/time of distribution; customer identifier; distribution media and prices; distribution address; mode of distribution; and volume distributed [Note: Only customer metadata needed for diagnosing problems, security or usage measuring will be kept to be mindful of public privacy.]

- *Access* including [BNLT] customer identification; current available funds in customer account; date/time/duration of access; mode of access; volume of information accessed; nature and volume of data requested; and mode of distribution. [Note: Only customer metadata needed for diagnosing problems, security or usage measuring will be kept to be mindful of public privacy.]

- *Customer Assistance* including [BNLT] inventory information; browse imagery; data summaries; management information summaries; orbital projections; data set availability; distribution media alternatives; data distribution pricing policies and system user's guides. Such assistance may take the form of interactive, electronic distribution of information coupled with human interaction in the form of a science and/or data help facility.

3.3 Archive Metadata

Information, which is associated with the data sets held within and distributed by the ADS, will be maintained. Such information will be stored in a form and fashion, which allows rapid, indexed access; data integrity; is presented in a human readable form; and will be an integral part of an ADS data management (DM) process. The following activities and information will be monitored, maintained, and made available for the life of the ADS:

- data and data product descriptions including [BNLT] data formats and structures; periods of observation; electronic signature and/or checksums, and periods of holding at the ADS

- data and product algorithms used by the IDPS and SDS

- known anomalies within data or data products

- all metadata will follow NOAA metadata standards (FGDC, ISO)

4.0 Ingest Management

4.1 Integrated Data Processing Segment (IDPS) to ADS

The ADS will ingest or receive operational data (RDRs, SDRs, EDRs and associated information) directly from the NESDIS/NCEP "Central" IDPS at Suitland, MD, through direct, dedicated telecommunications links on a 24x7 operational basis. IDPS-to-ADS links will be managed and maintained by the ADS to ensure continuous archive accessibility to data. Transfer of data is expected to be file transfer protocol (FTP push or pull) to protected ADS disk storage (for ingest).

The IDPS will be responsible for making data sets available and accessible. The IDPS and ADS system management processes will maintain "handshaking" protocols which will provide for the maximum amount of automation of the ingest element. Should a failure occur, the ADS system management function will alert ADS personnel through the ADS metadata management function that a problem has occurred and human intervention may be required. ADS operational personnel will be available for emergency ingest operations on an around-the-clock basis.

The Ingest element will confirm acceptance of data through agreed upon practices and procedures with the IDPS, possibly through file attribute comparisons (signatures) and/or "checksum" determination. Once a data set has been physically transferred (copied) and completeness has been confirmed, data will be considered "provisionally accepted" and metadata indicating that status will be generated. Upon "full acceptance" of data sets, (described below), IDPS responsibilities for these data sets can terminate. However, under the principles of good business practices, the IDPS will maintain all non-attached data sets at their facility for a sufficient enough time, (currently 24 hours is planned), to ensure the ADS can re-ingest, if necessary.

ADS provisionally-accepted data will be inventoried by the ingest element. Inventory information will be generated through the use of an IDPS-ADS data set naming convention (to be agreed upon by the data suppliers and the ADS) and through a preliminary, quality assurance, scan of each data set. This scan, or read, of data sets will provide information about completeness, quality, format, and structure. Errors or anomalies detected at this phase will require notification of both the IDPS and ADS system management functions and probable human intervention to rectify the situation. This interactive process will continue until a resolution is achieved. Once resolution is achieved, data sets will be "fully accepted", and those with errors/anomalies shall be so indicated in the ADS inventory.

Some data sets will be candidates for the creation of visualization or low-resolution browse products. If ingested data sets are scanned during the acceptance phase, browse creation can occur simultaneously, otherwise browse creation will be an independent process (it is understood that the SDS may also provide browse images). Storage of browse imagery will be the responsibility of the ingest element, while access and display will be the responsibility of the metadata management element. Systematic safeguards to reduce redundancy between SDS and IDPS browse imagery creation will be implemented.

The ingest element will allow for "standing order" or subscription distributions of data sets. Any data sets, which qualify for such distribution will be extracted at ingest for the subscriber and passed on to the distribution management element.

The ingest element for any data set terminates when the data achieves fully acceptable status (best quality that can be achieved from the original data stream). The physical location and data set name of acceptable data are passed to the data storage element.

4.2 Science Data Segment (SDS) to ADS

The ADS will ingest or receive climate data records and derived climate products directly from the NPP Science Data Segment (SDS), currently planned to be at NASA Goddard Space Flight Center in Greenbelt, MD. Any/all dedicated SDS-to-ADS links will not be considered "operational" links for NESDIS data management, thus will be managed and maintained by NASA; co-managed by NASA and NOAA under documented agreement; or managed by third parties, including commercial service providers. Transfer of data is expected to be FTP push or pull, to protected ADS on-line storage (for ingest).

The SDS will be responsible for making data sets available and accessible. The SDS and ADS system management processes will maintain "handshaking" protocols which will provide for the maximum amount of automation of the ingest element. Should a failure occur, the ADS system management function will alert ADS personnel through the ADS metadata management function that a problem has occurred and human intervention may be required. Since the SDS plans to retain the data sets they create, ingest from the SDS will not be an operational process, and the ADS operational personnel need not be available for emergency ingest operations on an around-the-clock basis.

The Ingest element will confirm acceptance of data through agreed upon practices and procedures with the SDS, possibly through file attribute comparisons (signatures) or checksum determination. Once a data set has been physically transferred/copied and completeness confirmed, data will be considered "provisionally accepted" and metadata indicating that status will be generated. Upon "full acceptance" of data sets (described below), SDS archive and distribution responsibilities for these data sets can terminate, if desired. However, under the principles of good business practices, the SDS will maintain all non-attached data sets at their facility for sufficient enough time to ensure the ADS can re-ingest, if necessary.

ADS provisionally-accepted data will be inventoried by the ingest element. Inventory information will be generated through the use of an SDS-ADS data set naming convention and through a preliminary, quality assurance, scan of each data set. This scan, or read, of data sets will provide information about completeness, quality, format, and structure. Errors or anomalies detected at this phase will require notification of both the SDS and ADS system management functions and probable human intervention to rectify the situation. This interactive process will continue until a resolution is achieved. Once resolution is achieved, data sets will be "fully accepted," and those with errors/anomalies shall be so indicated in the ADS inventory.

Some data sets will be candidates for the creation of visualization or low-resolution browse products. If ingested data sets are scanned during the acceptance phase, browse creation can occur simultaneously, otherwise browse creation will be an independent process (it is understood that the SDS may also provide browse images). Storage of browse imagery will be the responsibility of the ingest element, while access and display will be the responsibility of the metadata management element. Systematic safeguards to reduce redundancy between SDS and IDPS browse imagery creation will be

implemented.

The ingest element will allow for "standing order" or subscription distributions of data sets. Any data sets, which qualify for such distribution will be extracted at ingest for the subscriber and passed on to the distribution management element.

The ingest element for any data set terminates when the data achieves fully acceptable status (best quality that can be achieved from the original data stream). The physical location and data set name of acceptable data are passed to the data storage element.

5.0 Data and Product Archive/Preservation and Migration Management

All data and information received will be transcribed to long-term storage media. All processes for the transcription and maintenance of this data (to include quality electronic signatures/checksums), and information will be controlled by a Commercial Off The Shelf (COTS) storage management system (SMS) and will be performed on an automated basis. The SMS will track data disposition and provide such tracking information as metadata to be shared by other system elements, including data retrieval and customer interface. The physical location and media types used will be transparent to the general public. Data quality and integrity will be insured as a function of the SMS.

Data used for servicing customer requests will be stored using a hierarchy of storage management regardless of physical media. The SMS will maintain tracking information about data (including quality information), which is available on any media and will provide such information to the retrieval element for efficient servicing. This storage hierarchy will not preclude long term storage, described above, but will supplement it.

Data will also be provided under the direction of the SMS for storage at an off-site, disaster backup facility. Such data will maintain the form and structure of the permanently stored ADS data and is subject to retrieval (copied from the off-site storage) only at times of emergency. The SMS will provide information to the system element that will manage the transmission of data to the backup facility on a routine, scheduled basis.

Data archived on computer compatible media will be migrated to preclude media obsolescence. Data migration or transcription to new media will be performed on a scheduled basis. All migrations will perform quality checks for data integrity and will be transparent to ADS customers. Migrations will be controlled by the SMS; will be performed at the discretion of NESDIS management; as time and resources permit; and as storage technologies improve in speed of I/O and media density.

6.0 Data and Product Retrieval and Distribution Management

All data and information retrievals will be managed by the COTS SMS (see 5.0). The customer interface will control what data and information is requested (for either electronic or some form of media delivery) and the retrieval portion of the SMS will locate and extract that data. With the hierarchical approach, such retrieved data may be found regardless of physical media.

The retrieval element will also provide all mechanisms for sub-setting or super-setting data based on ADS customer demand, including the Climate Analysis and Research System (CARS), Distributed CDR Algorithm Validation System (DCAVS), Climate Data Management System (CDMS) of the SDS, and service levels provided. Sub-setted and super-setted data will require processing and temporary storage on-line, prior to distribution. The system element will control the disposition of such "new" data sets. In the case where the customer requests the data electronically, the retrieved data set(s) will be staged to a secure FTP site for pick-up.

The retrieval element will also provide for the transcription of data to computer compatible media for physical shipment to customers. Again, the system element will manage the transcription of data to transportable media subsequent to extraction.

All bulk data retrievals for migration will be managed by the SMS. For media migration, the process will be automated such that data will be extracted from existing/old media and transcribed to new media following proven quality assurance practices. The location of data, whether on old or new media, will remain transparent to the customer interface during any migration process, i.e., data extraction will be performed on a transparent basis during migration(s).

Where appropriate, primarily for derived data products, the retrieval element will provide the mechanism for reformatting data per customer request. Such reformatting may include the translation to another format other than the one archived, (e.g., HDF or NetCDF), for customer analysis and further processing. As with data sub-sets/super-sets, all reformatted data will be stored temporarily on on-line. Location will be controlled by the system element.

The primary mode of ADS data delivery will be through electronic distribution. The ADS will utilize FTP (or subsequent IT standard) protocols to provide data. All FTP areas will exist outside of the ADS firewall. For interactive customers, requested data will be made available to pull from an FTP area. For standing orders or subscription customers such as the SDS CARS, DCAVS and CDMS, data will be provided as either pull or push based upon agreement.

Bulk data retrieved for delivery to the SDS will be either pushed or pulled via the established telecommunications link(s) between the ADS and the SDS (see 4.2) based on mutual agreement.

Data will also be provided on computer compatible media for off-line delivery to customers. Such data will be distributed per standard NESDIS practices and procedures

established by NESDIS management.

7.0 Customer Interface Management

The ADS will provide a single, interactive customer interface available via the Internet, or its subsequent replacement. This interface will reside outside of the ADS system and will provide a suite of customer services which include access to archive metadata, including inventories, data and data set descriptions, browse imagery, and all data delivery options. This element will fabricate all commands passed to the retrieval element, including sub-setting/super-setting and reformatting criteria requested.

The customer interface will also collect required information regarding appropriate cost recovery. It will interface with an e-commerce cost-servicing element, for billing and accounting as appropriate. The cost-servicing element will interact with the customer and provide the ADS customer interface with the authorization to fulfill the pending request. The ADS will not provide a unique cost transaction interface.

The ADS will provide access to the publicly available areas via interfaces with several existing and potential NOAA, other government, and academic networks, including, but not limited to NOAA Server and NOAA National Data Center (NNDC) Servers.

8.0 Data Reprocessing

There is currently no requirement for data reprocessing for the ADS segment.